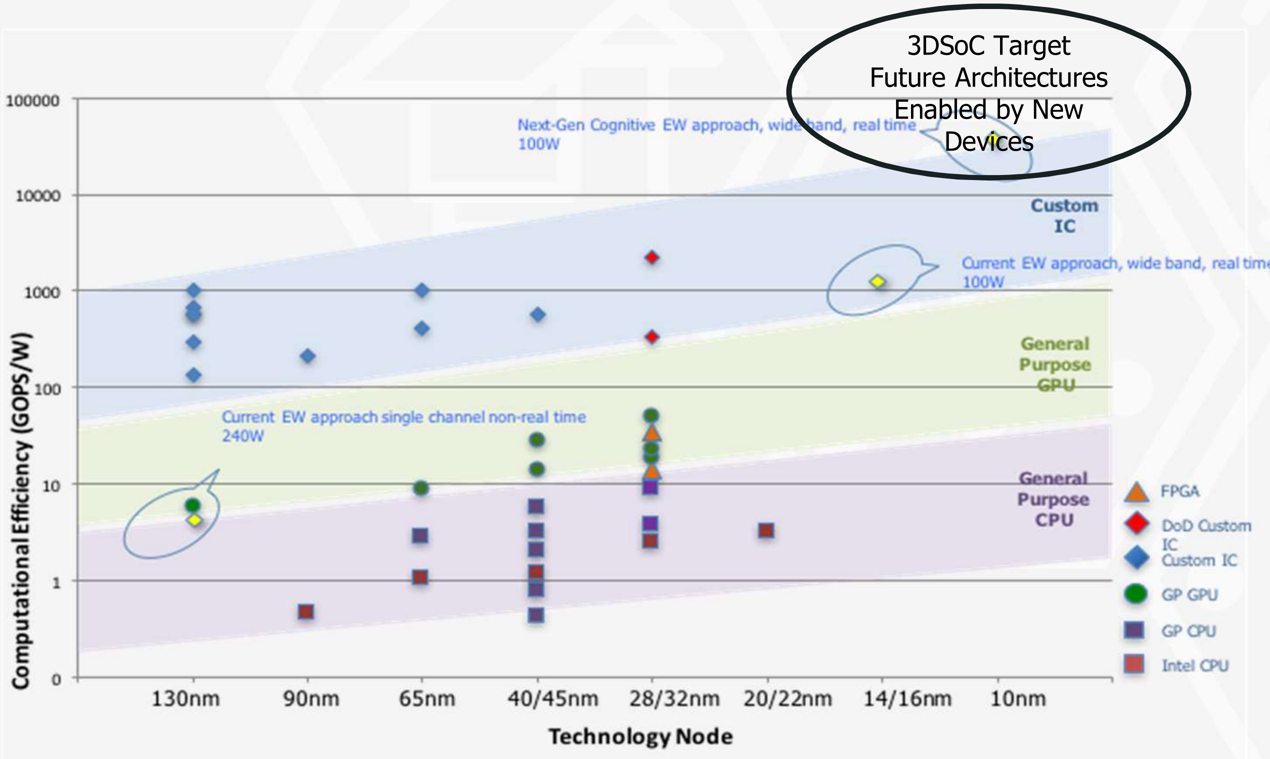
THE ELECTRONICS RESURGENCE INITIATIVE

LINTON SALMON

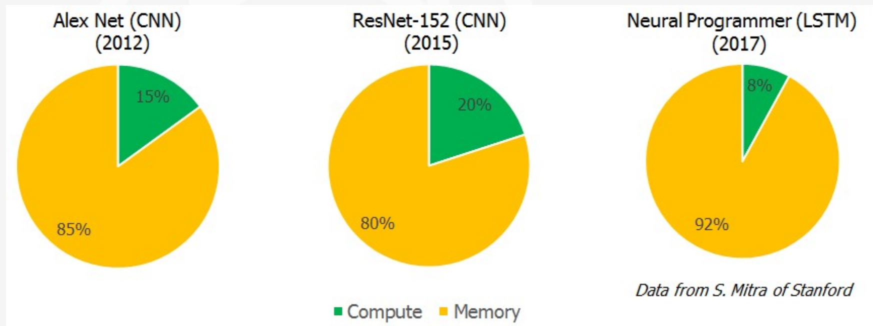
PROGRAM MANAGER
DARPA/MTO

3 DIMENSIONAL MONOLITHIC SYSTEM ON A CHIP (3DSOC)

BEYOND 2D-SCALING TECHNOLOGY IS CRITICAL FOR DOD



THE MEMORY BOTTLENECK



Note: These data representative of systems that benefit from massive caching, parallelism, and pipelining

Amdahl's Law

$$\text{Overall Speedup} = \frac{1}{(I-F)+\frac{F}{s}}$$

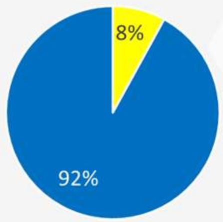
F = Fraction enhanced
s = Speedup of enhanced fraction

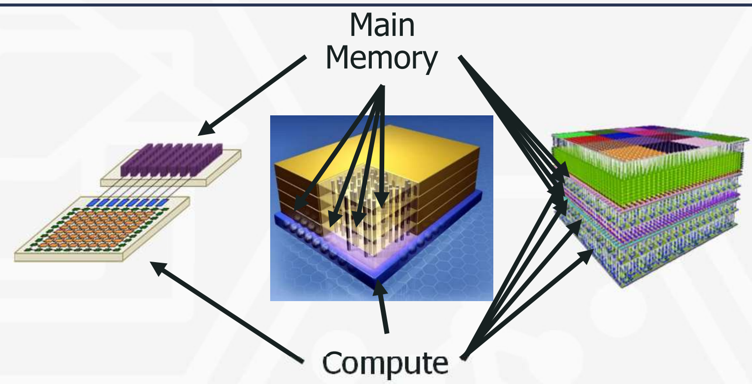
Compute Speedup is throttled by memory since memory access is not speeding up

Memory Speedup	Compute Speedup	Fraction Compute	System Speedup
1X	100X	8%	9%
10X	1X	8%	580%

Arbitrary increases in speedup of computation have limited impact on system performance unless the memory bottleneck is addressed

ADDRESSING THE MEMORY LIMITATION





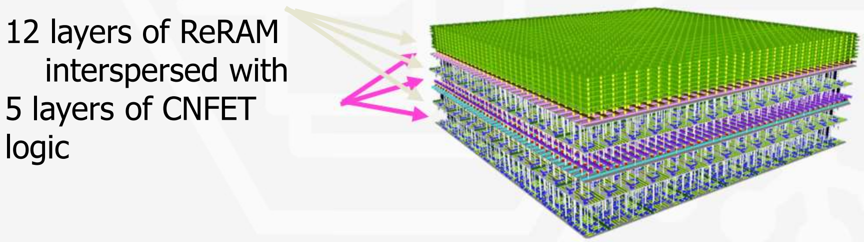
Production 2D	Development 3D TSV Package	DARPA 3DSoC

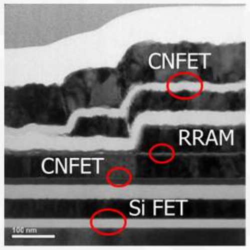
Memory Access Parameter	2D		3D TSV Package		3DSoC
Total I/O	512		1K		33K
Max Bandwidth (Gb/s)	400		1K		46K
Memory access energy (pJ/bit)	52		32		1.5
VDD (Volts System)	1.6		1.2		0.6

3DSoC increases the IO count and bandwidth by >50X from current 2D fabrication architectures

AN INTEGRATED, MONOLITHIC SOC (3DSOC) SOLUTION

An integrated flow that fabricates 3D logic and memory on a single die





from S. Mitra of Stanford University

Critical characteristics for a monolithic solution

- Must permit new architectures that leverage fast, configurable access to non-volatile main memory
- Stackable 3D logic and memory functions that allow new architectures
  - Low temperature formation
  - Logic AND memory
  - High density of memory – at least 4GB (Giga-Byte)/die
- Possible to fabricate in existing domestic, commercial, high-yielding infrastructure
  - 90nm on 200mm wafers
  - High yield on large SoCs

SIMULATION RESULTS FOR MACHINE LEARNING

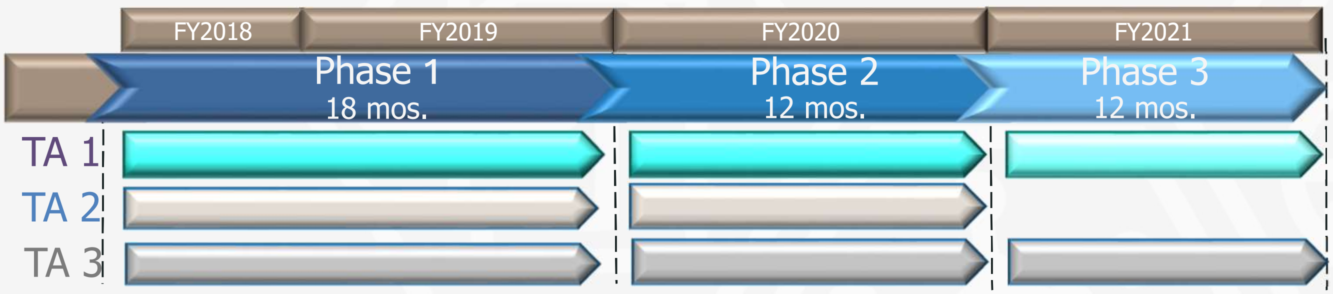
LSTM Network	Model Size	Training/Inference	2D at 7nm	
			Benefit 3DSoC at 7nm	Benefit 3DSoC at 90nm
Language Model	2.5 Gbytes	Training	645X	75X
		Inference	626X	73X
Neural Programmer	1 Gbyte	Training	359X	40X
		Inference	493X	55X
Image Captioning	150MByte	Training	367X	41X
		Inference	323X	35X

from S. Mitra of Stanford University

- 2D vs 3DSoC comparison
  - 2D: 7nm technology for accelerator and 4GB of off-chip DRAM main memory
  - 3DSoC: 90nm Carbon Nanotube FET (CNFET) for accelerator and 4GB of on-chip ReRAM (non-volatile) memory
- Uses published traces from an accelerator SoC and the LSTM algorithm
- Benefit = $(E*t)_{2D} / (E*t)_{3D}$ **
- Benefits would be enhanced if design targeted at 3DSoC technology

** E=Energy of computation t=execution time of computation

3DSOC PROGRAM SCHEDULE



Phase 1 Outcomes
- Initial 3DSoC process defined
- PDK V0.5 defined
- 3DSoC technology benefits simulated
- First pass DEC fabricated and tested
- Initial 3DSoC EDA tools released

Phase 2 Outcomes
- 3DSoC process demonstrated
- PDK V1.0 released for design
- 3DSoC benefits demonstrated
- Final DEC design fabricated
- 3DSoC EDA tools released for targeted designs

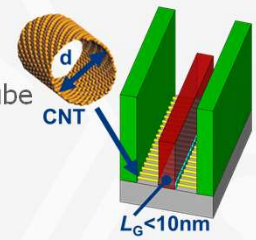
Phase 3 Outcomes
- 3DSoC process used for external designs
- Final PDK released for design
- MPW runs successfully yielded
- 3DSoC EDA tools released for general designs

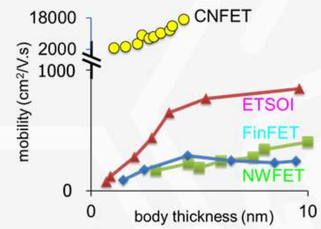
- TA-1: Developing the 3DSoC fabrication process
  - Establish unit processes and flow integration
  - Define the 3DSoC technology PDK
- TA-2: Designing and Implementing the DEC
  - Design 1st and 2nd pass DEC design
  - Foster use of the DEC to drive development and yield
- TA-3: Developing the 3DSoC EDA design flow
  - Develop EDA tools for 3DSoC compute/memory designs
  - Support tools for advanced 3DSoC designs

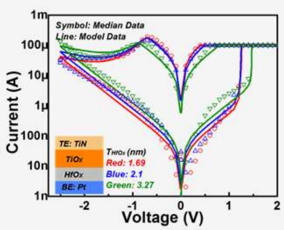
Metric	Goal
3DSoC Capability	> 50X 7nm 2D PaP
Hardware Accuracy	< 2% deviation from3DSoC technology targets
Yield	> 30% for full 3DSoC designs
EDA Tools	Successful use of EDA flow for a > 500M gate/4GB memory design

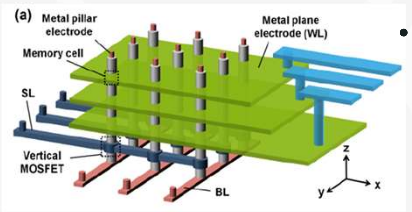
MIT/SKYWATER/STANFORD (3DSOC TA-1/TA-2)

- Carbon Nanotube transistor (CNFET) - ***MIT***
  - Current channel defined by 1.2nm Carbon Nanotube
  - Much higher channel mobility than CMOS devices
  - As a result, higher performance
  - Low fabrication temperature (<450C)





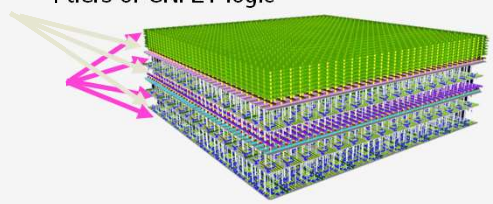


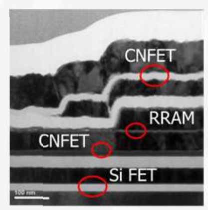


- Resistance RAM (RRAM) - ***Stanford***
  - Non-volatile memory (NVM)
  - High-density and scalable
  - Low fabrication temperature (<450C)

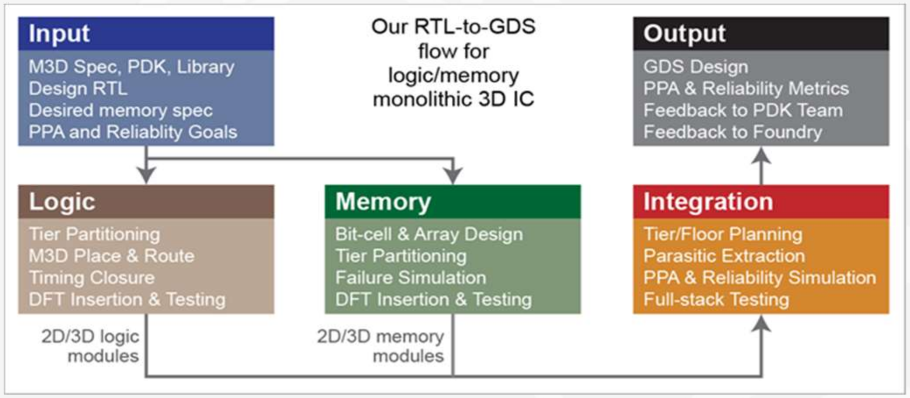
- 3D Integration – ***Skywater Tech Foundry***
  - Interspersed logic and memory layers
  - Interconnection pitch small (<300nm)
  - 4GB of on-die memory possible
  - Low fabrication temperature (<450C)

12 tiers of RRAM interspersed with 4 tiers of CNFET logic





GA TECH 3DSOC DESIGN SOFTWARE DEVELOPMENT (TA-3)



Characteristics of the Georgia Tech design tool approach

- Uses tier partitioning and subsequent integration to fully utilize 3DSoC technology
  - Partitions memory and logic into tiers based on design/technology defined characteristics
  - Partitions can be either large or small and can be interspersed
- Performs moderated place & route within and across device tiers
- Extracts all key in-tier and between-tier design parameters (resistance, capacitance, etc.)
- Provides a robust EDA environment to simulate 3DSoC performance and power
  - Augments existing 2D EDA tools as required
- Addresses key non-design activities
  - Reliability
  - Design for Test (DFT)

ERI ELECTRONICS RESURGENCE INITIATIVE

SUMMIT

2018 | SAN FRANCISCO, CA | **JULY 23-25**

MAX
SHULAKER

MIT

3DSOC: TRANSFORMING IDEAS INTO REALITY

3DSOC TEAM

Max Shulaker

Anantha Chandrakasan

Subhasish Mitra

H.-S. Philip Wong

Simon S. Wong

Brad Ferguson

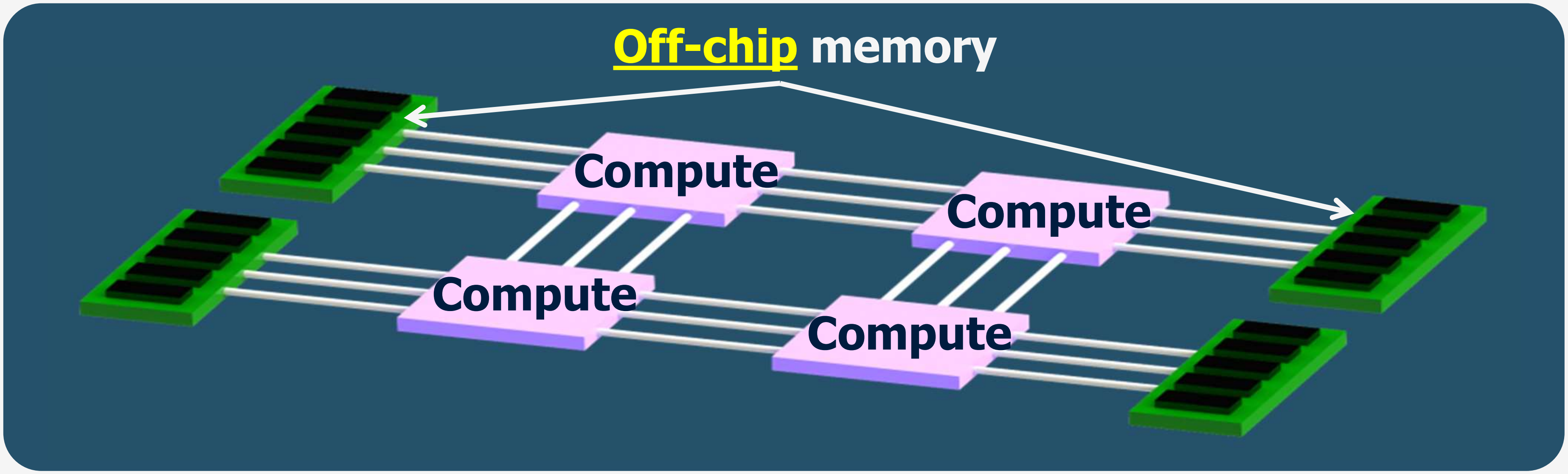
Mark Nelson

Jefford Humes

- This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA)
- The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government

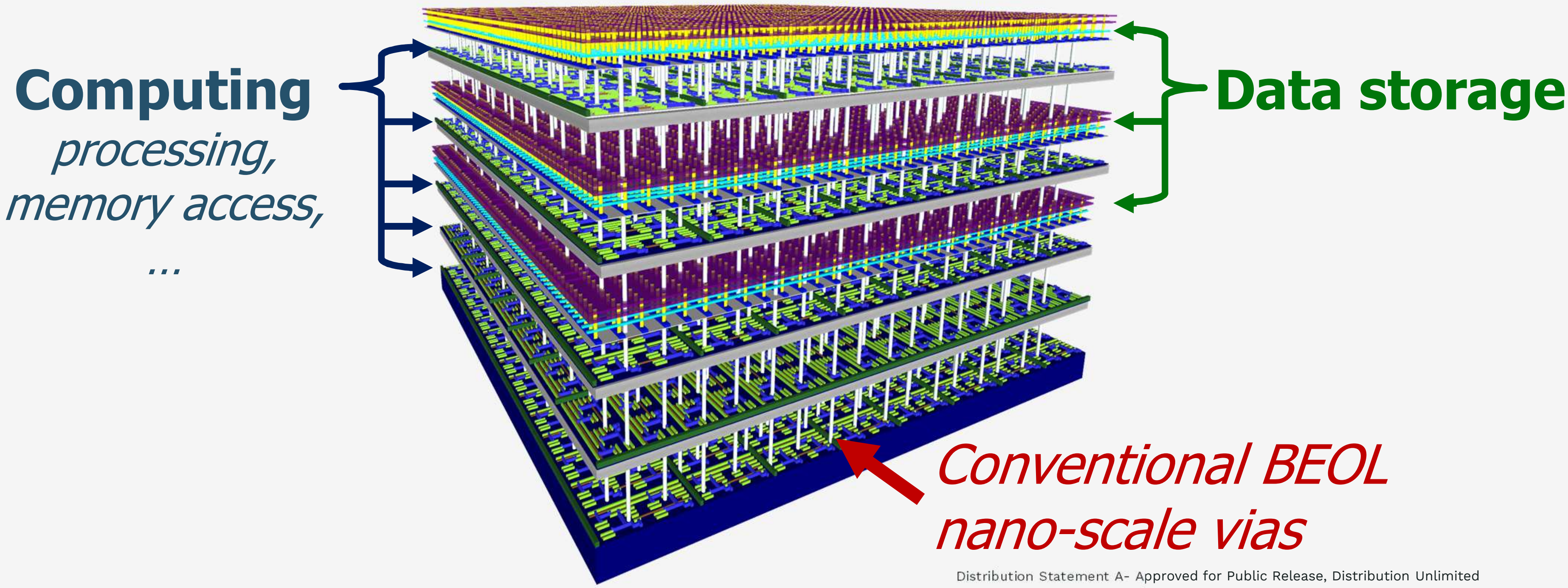
3DSOC: OUR APPROACH

- Transform:
  - *Conventional 2D computing system*



3DSOC: OUR APPROACH

- Monolithic 3D Integration
  - *Fine-grained integration: logic + memory*



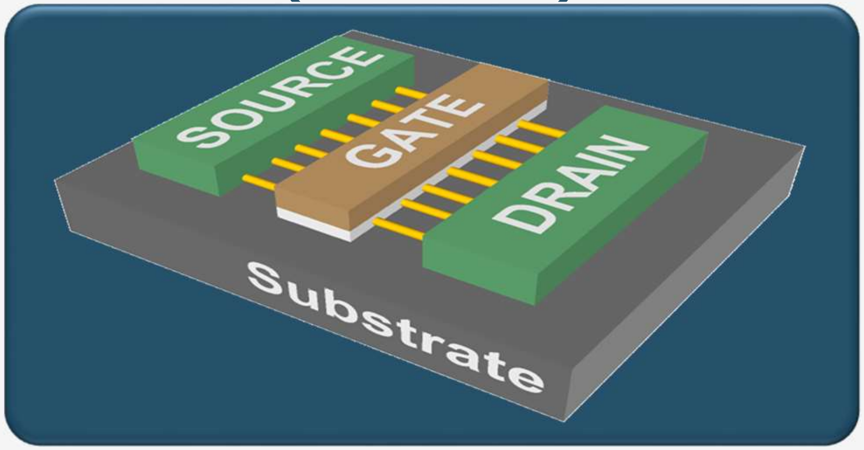
ENABLING TECHNOLOGIES

- Requires low temperature fabrication
  - *Challenging with conventional silicon CMOS*

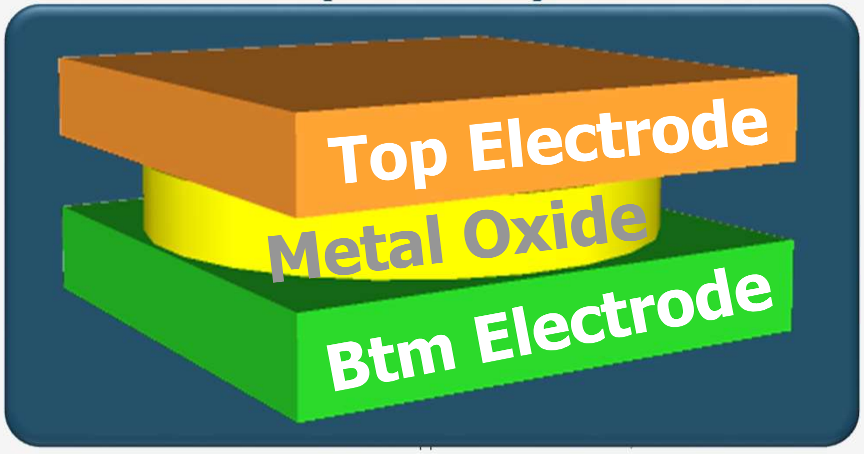
ENABLING TECHNOLOGIES

- Requires low temperature fabrication
  - *Challenging with conventional silicon CMOS*

Carbon Nanotube FETs
(CNFETs)



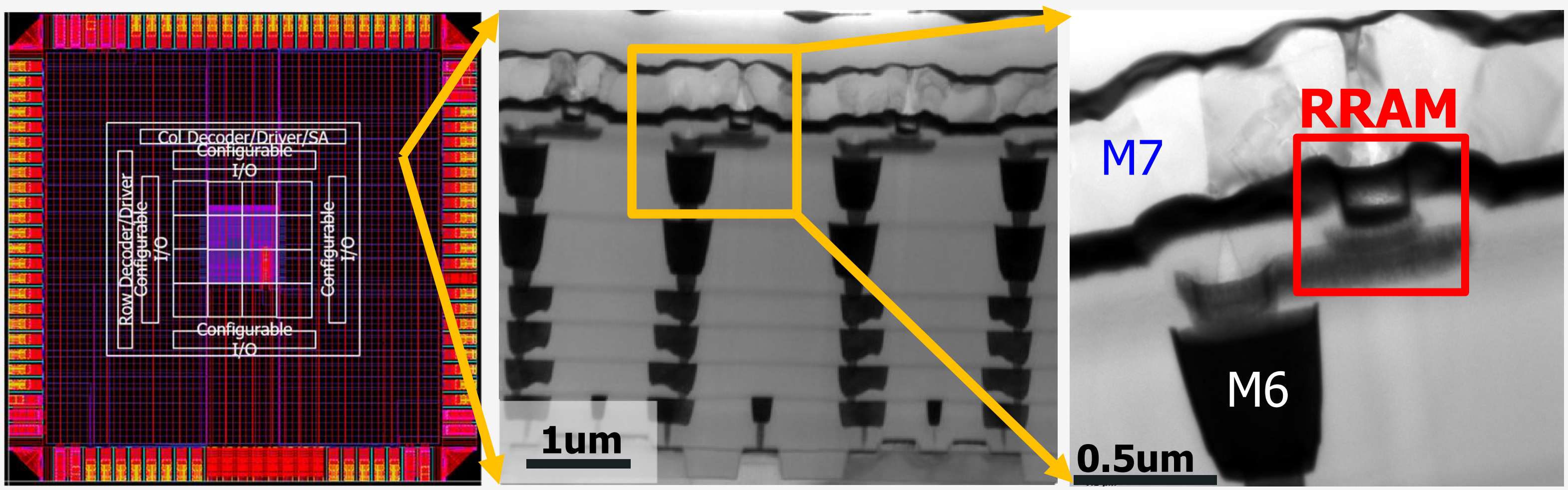
Resistive RAM
(RRAM)



RRAM TECHNOLOGY

- Dense on-chip non-volatile memory
  - *BEOL compatible*
  - *Ti/HfO$_X$/TiN stack*

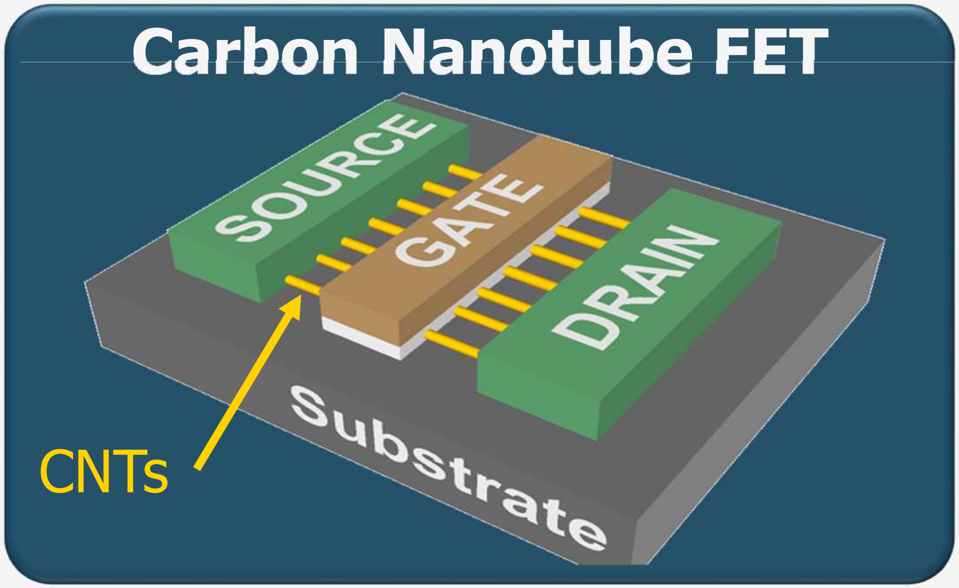
BEOL RRAM integration



source: Foundry/ Stanford

CNFET TECHNOLOGY

- Energy-efficient logic
  - **E**nergy **D**elay **P**roduct: ***10X*** *benefit*
  - *BEOL compatible*
  - *Relaxed 90 nm technology node*

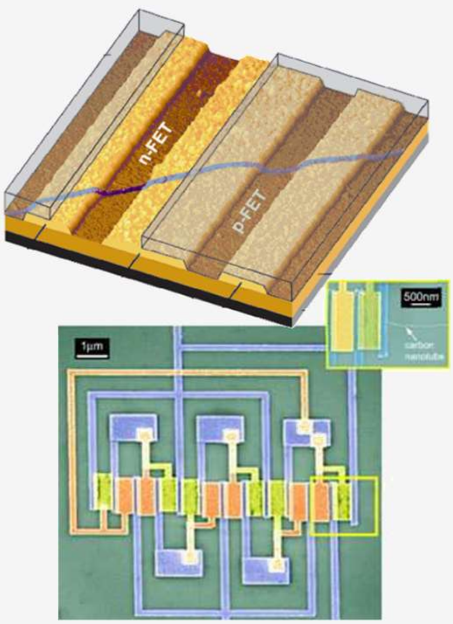


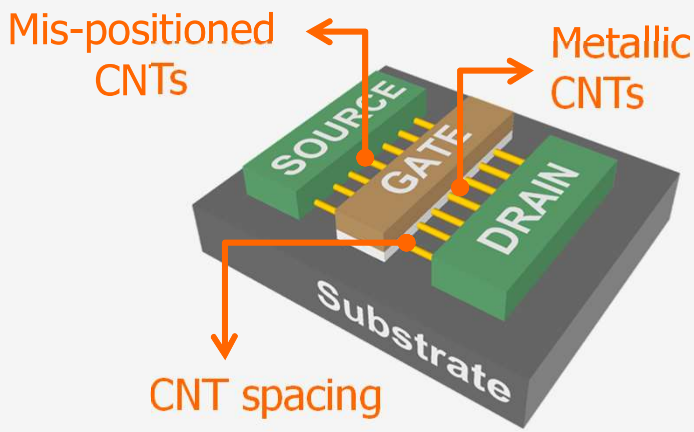
CNFETS: FROM DEVICES TO SYSTEMS

First CNFETs

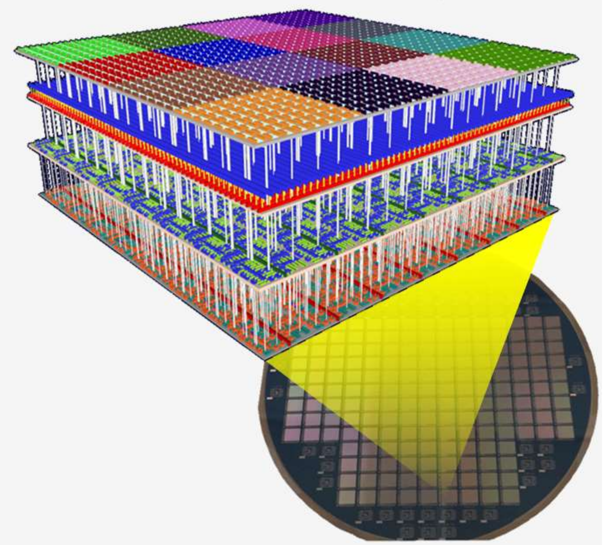
Overcoming Imperfections

Working Systems





VLSI compatible: processing + design solutions



Source: IBM

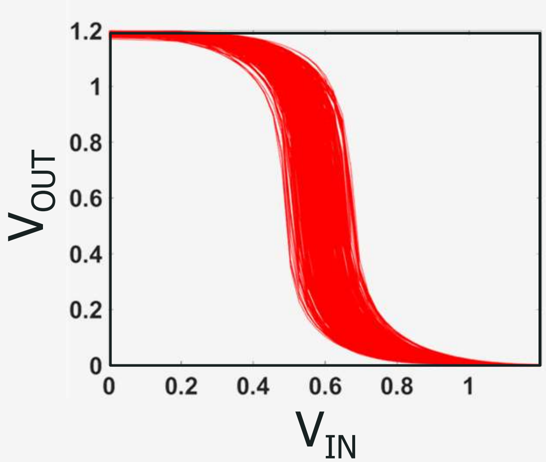
Source: Stanford

Shulaker 17

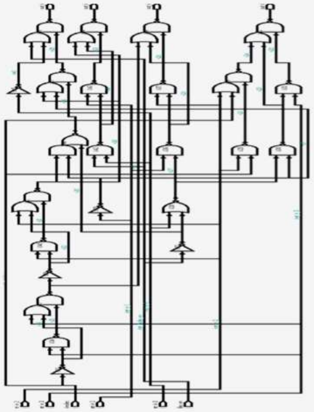
CNFETS: FROM LAB TO MANUFACTURING

materials | processing | design | manufacturing

Si-compatible; >99.99% purity



Robust doping; CNFET CMOS

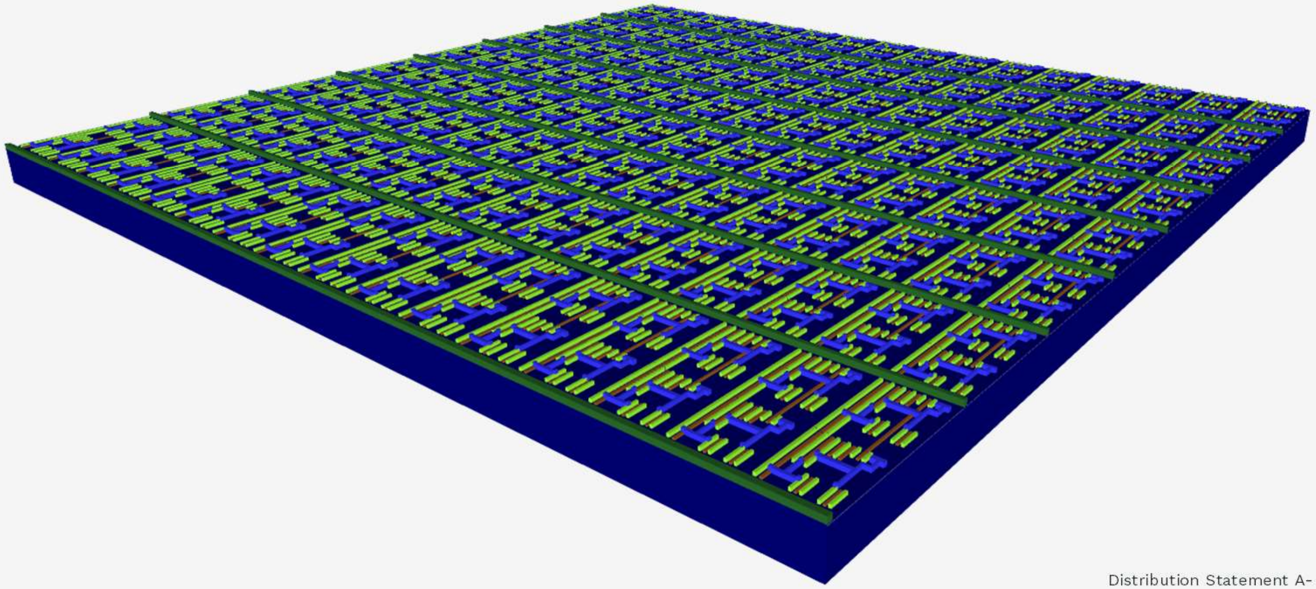


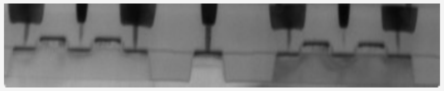
Immune to metallic CNTs

Commercial facilities

BUILDING THE 3DSOC

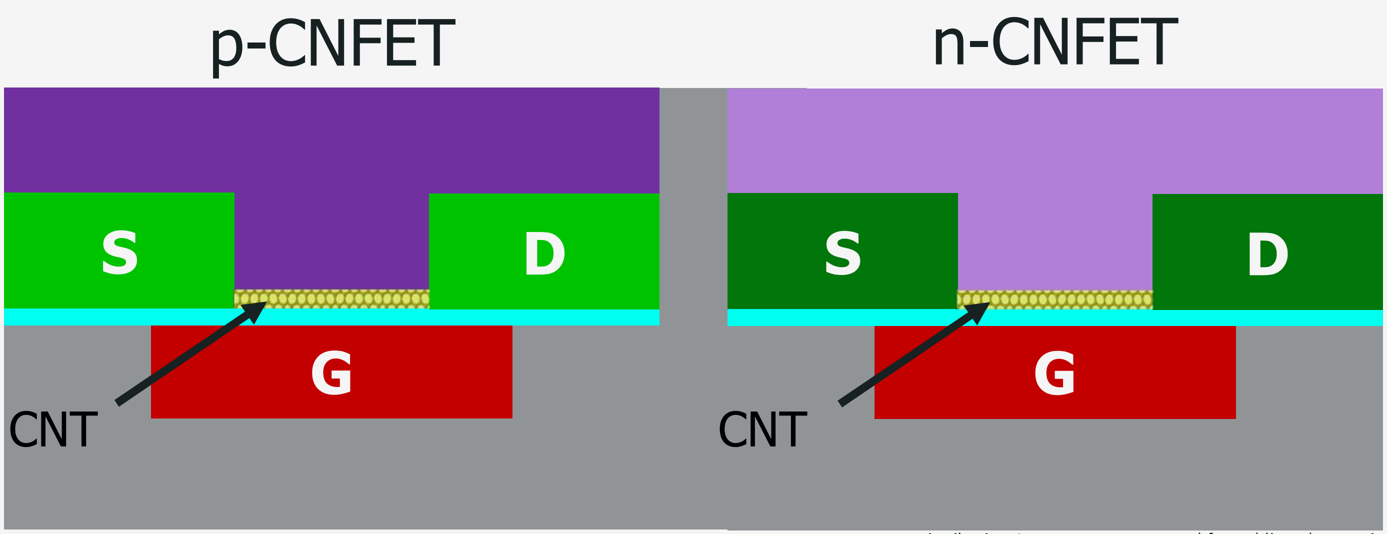
- Layer 1: CNFET CMOS logic
  - *compute*



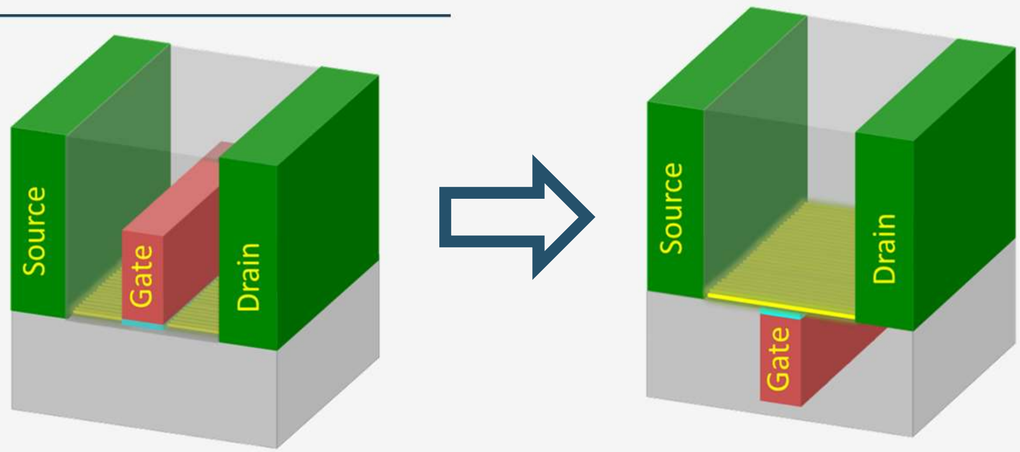


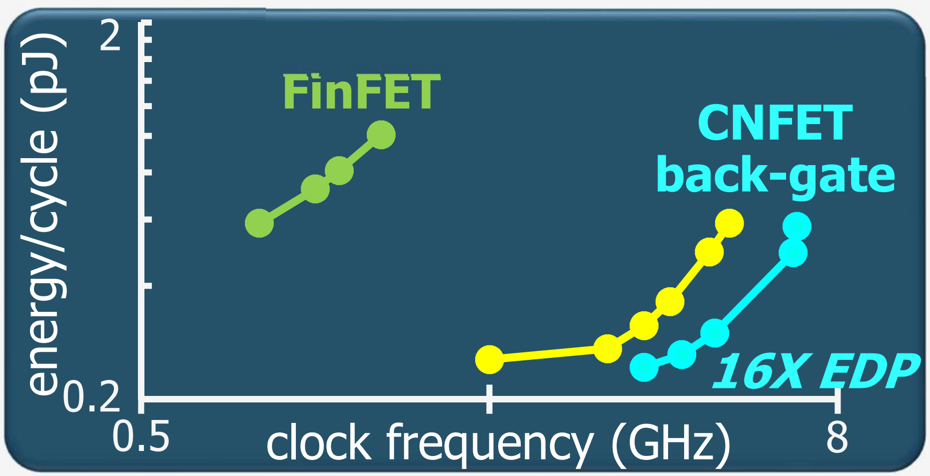
CNFET FABRICATION

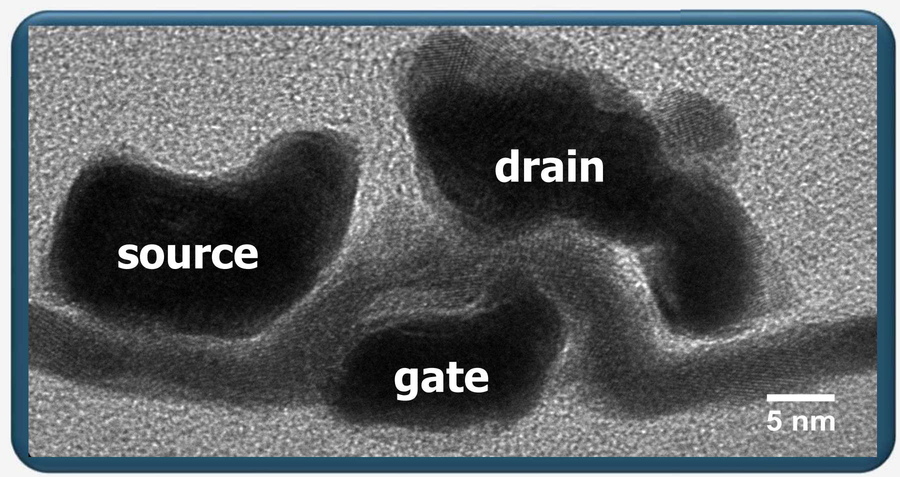
- Implementation:
  - *Complementary: CNFET-CMOS*
  - *Silicon-CMOS compatible*



CNFET FABRICATION

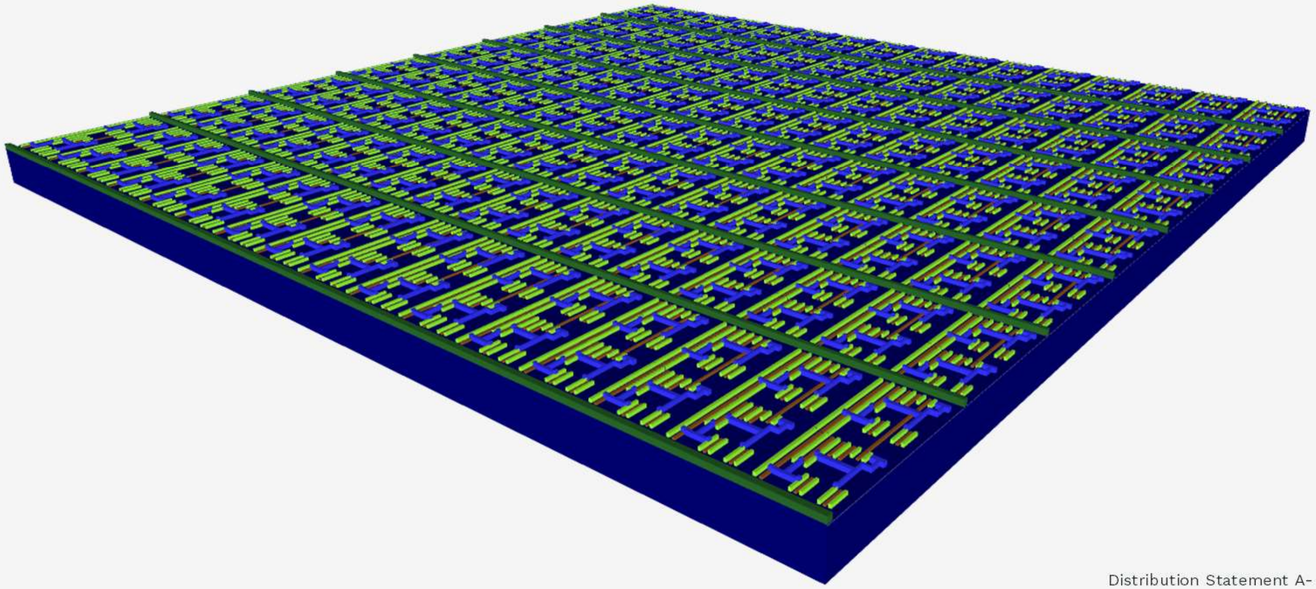


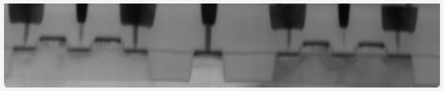




BUILDING THE 3DSOC

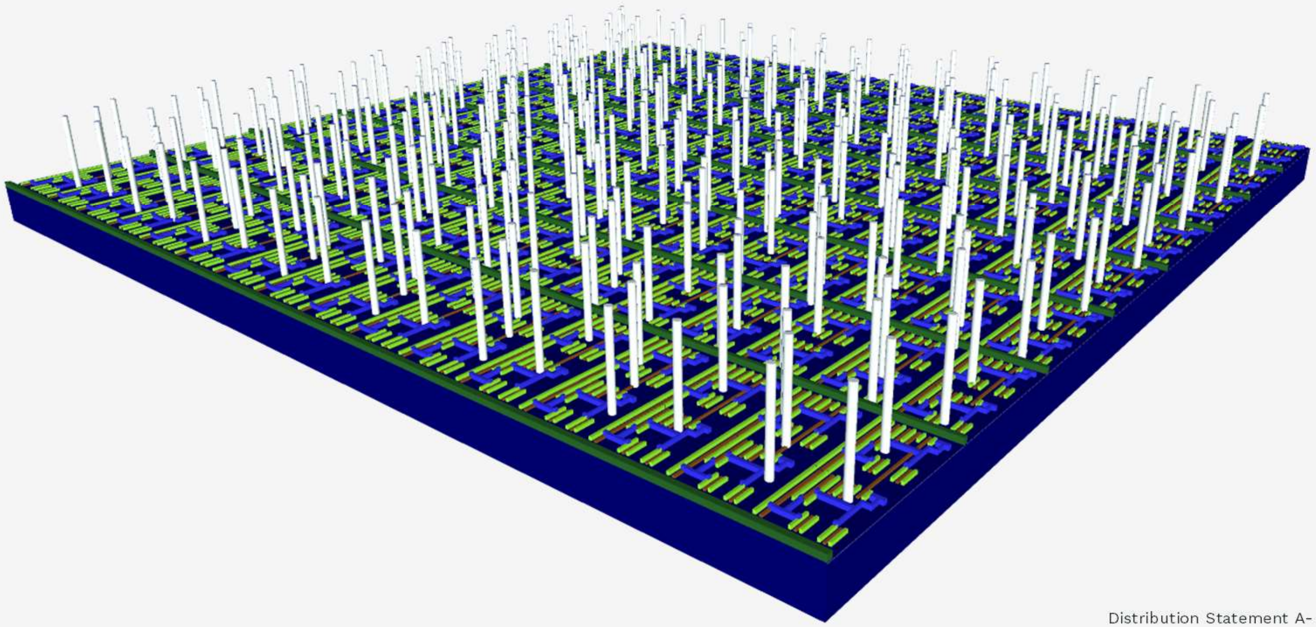
- Layer 1: CNFET CMOS logic
  - *compute*

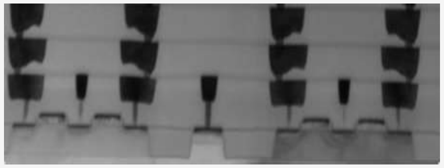




BUILDING THE 3DSOC

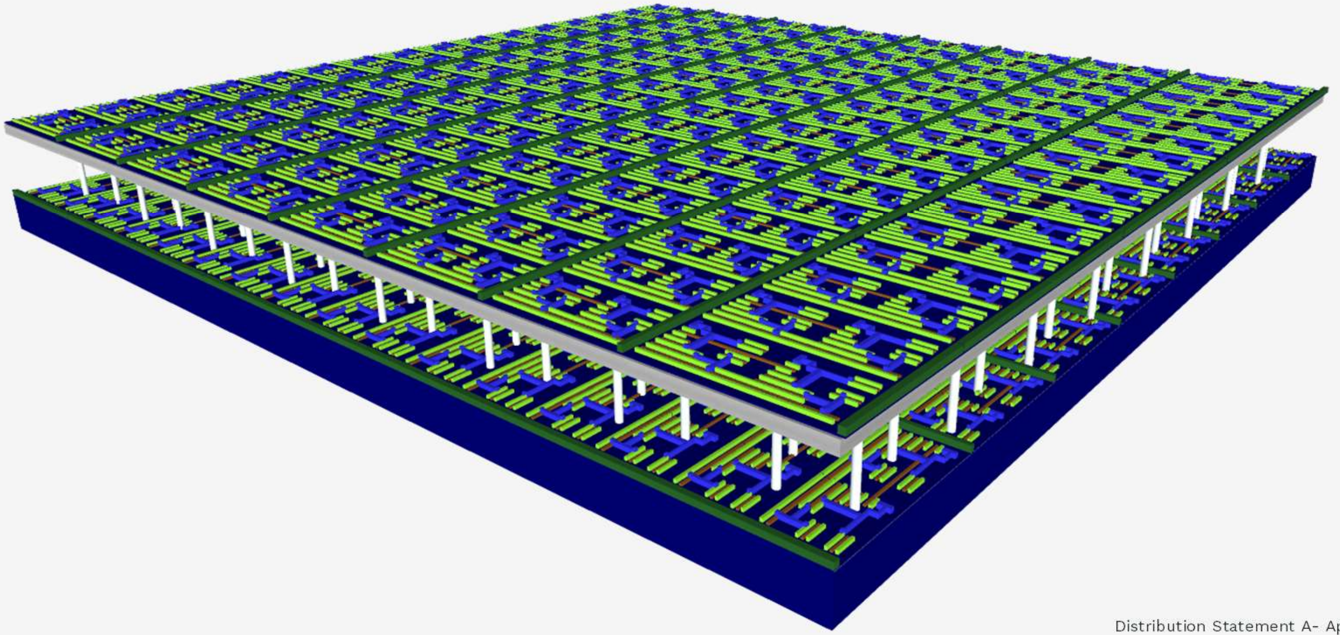
- Conventional BEOL nano-scale vias
  - *dual-damascene process*

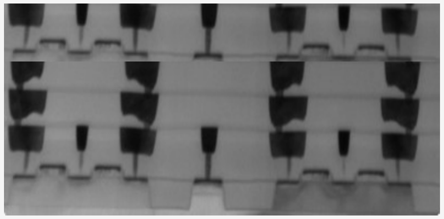




BUILDING THE 3DSOC

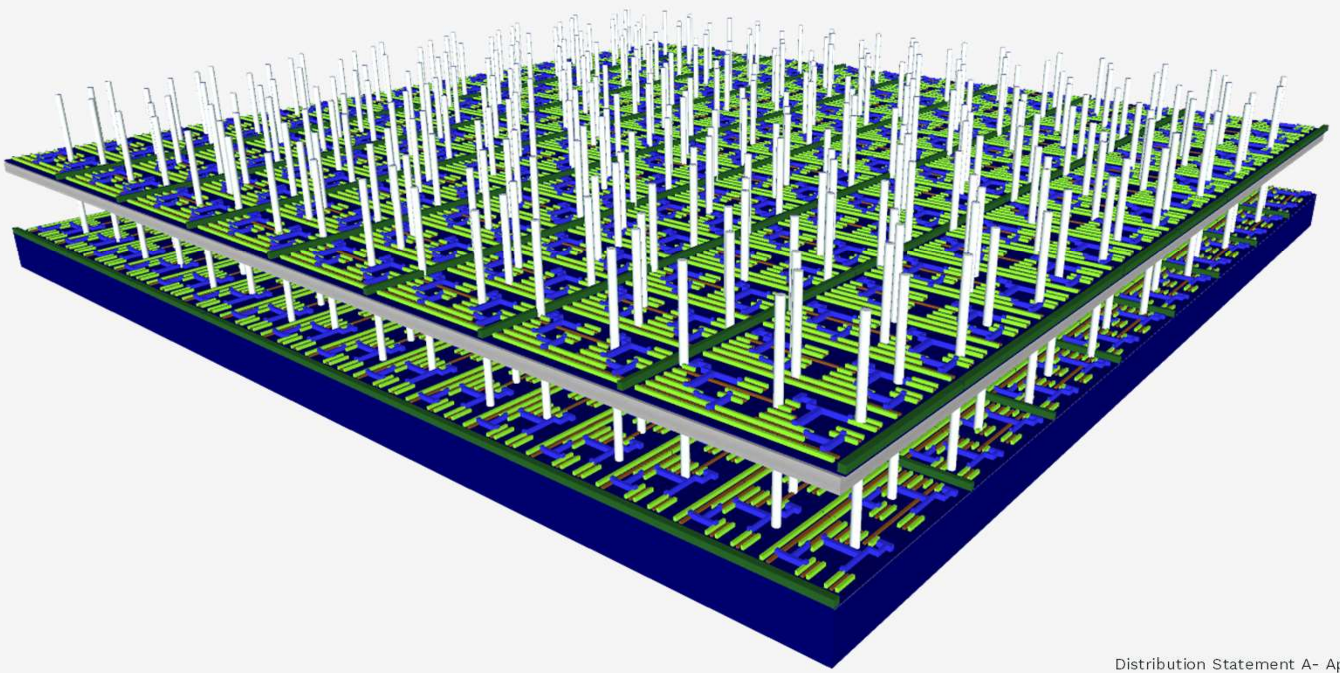
- Layer 2: CNFET CMOS logic
  - *compute*

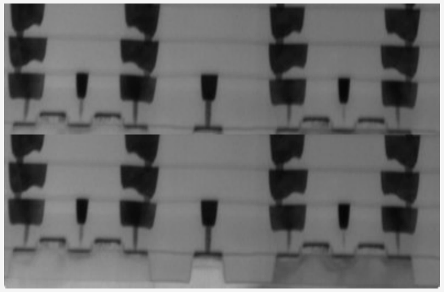




BUILDING THE 3DSOC

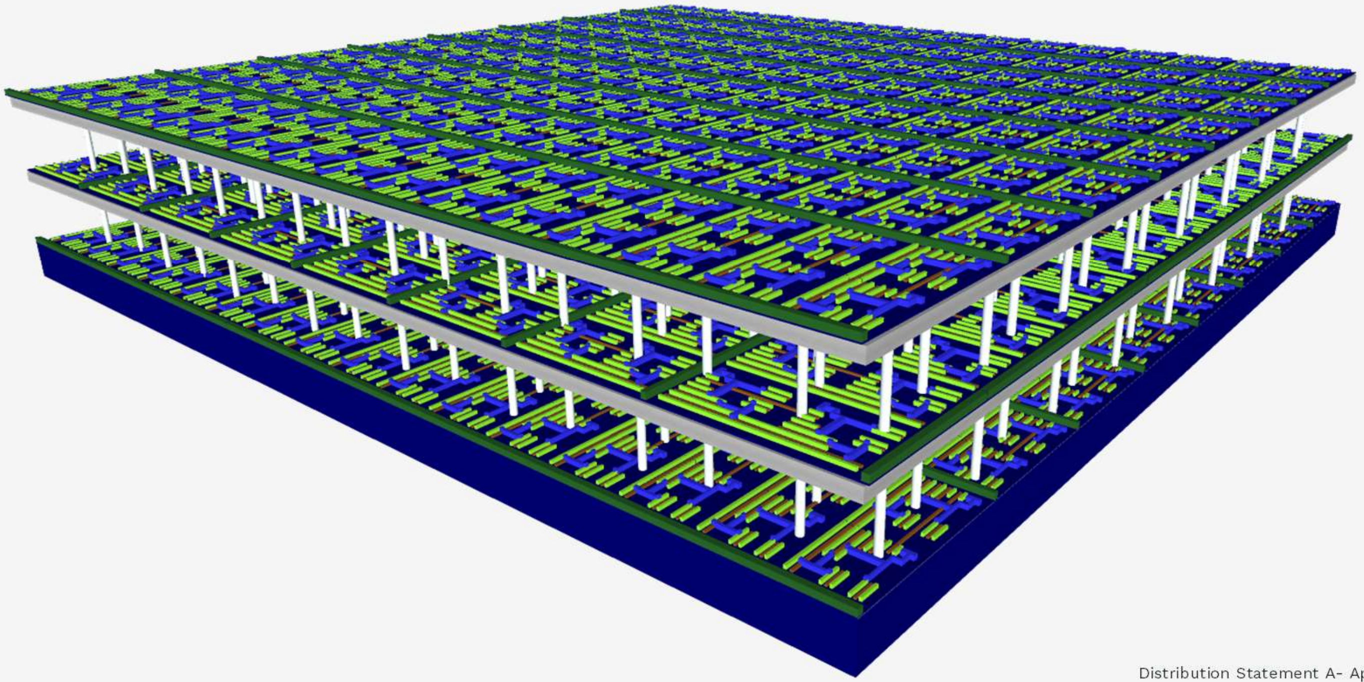
- Conventional BEOL nano-scale vias

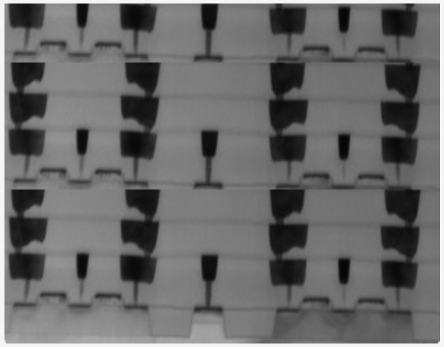




BUILDING THE 3DSOC

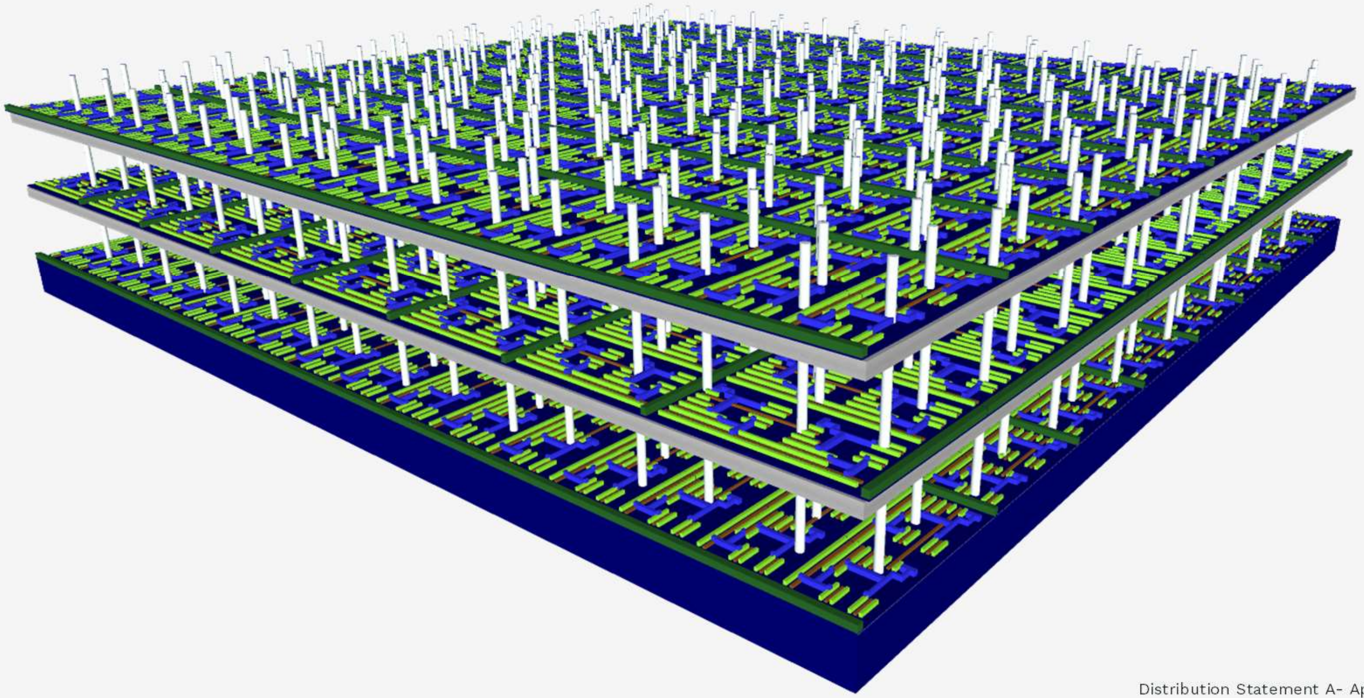
- Layer 3: CNFET CMOS logic
  - *memory access*

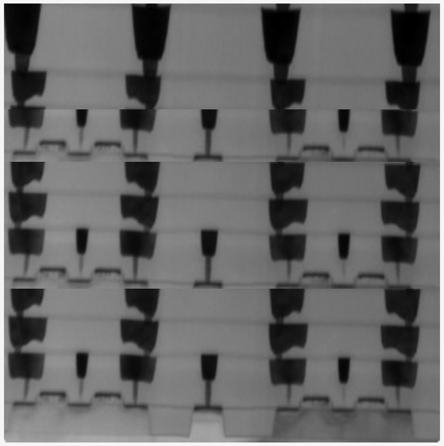




BUILDING THE 3DSOC

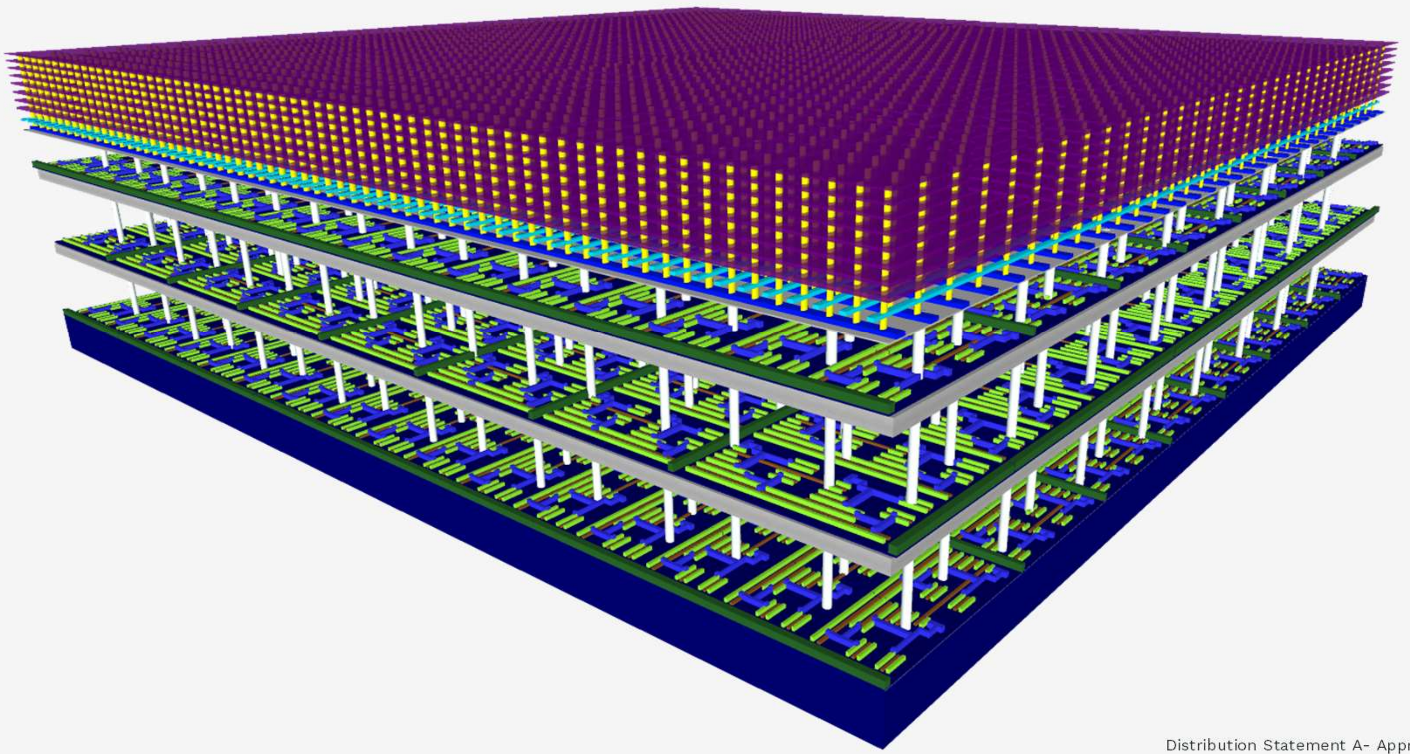
- Conventional BEOL nano-scale vias

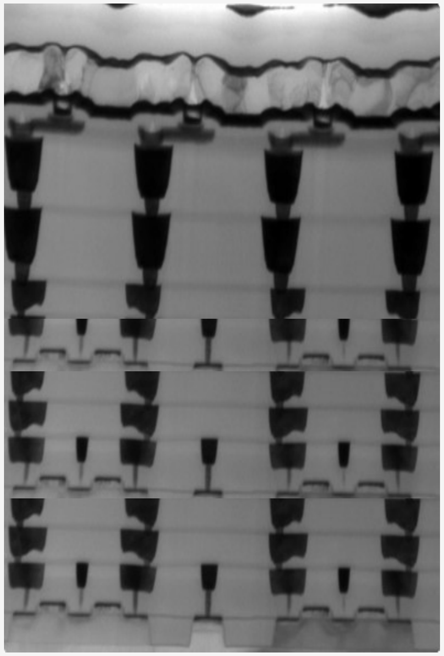




BUILDING THE 3DSOC

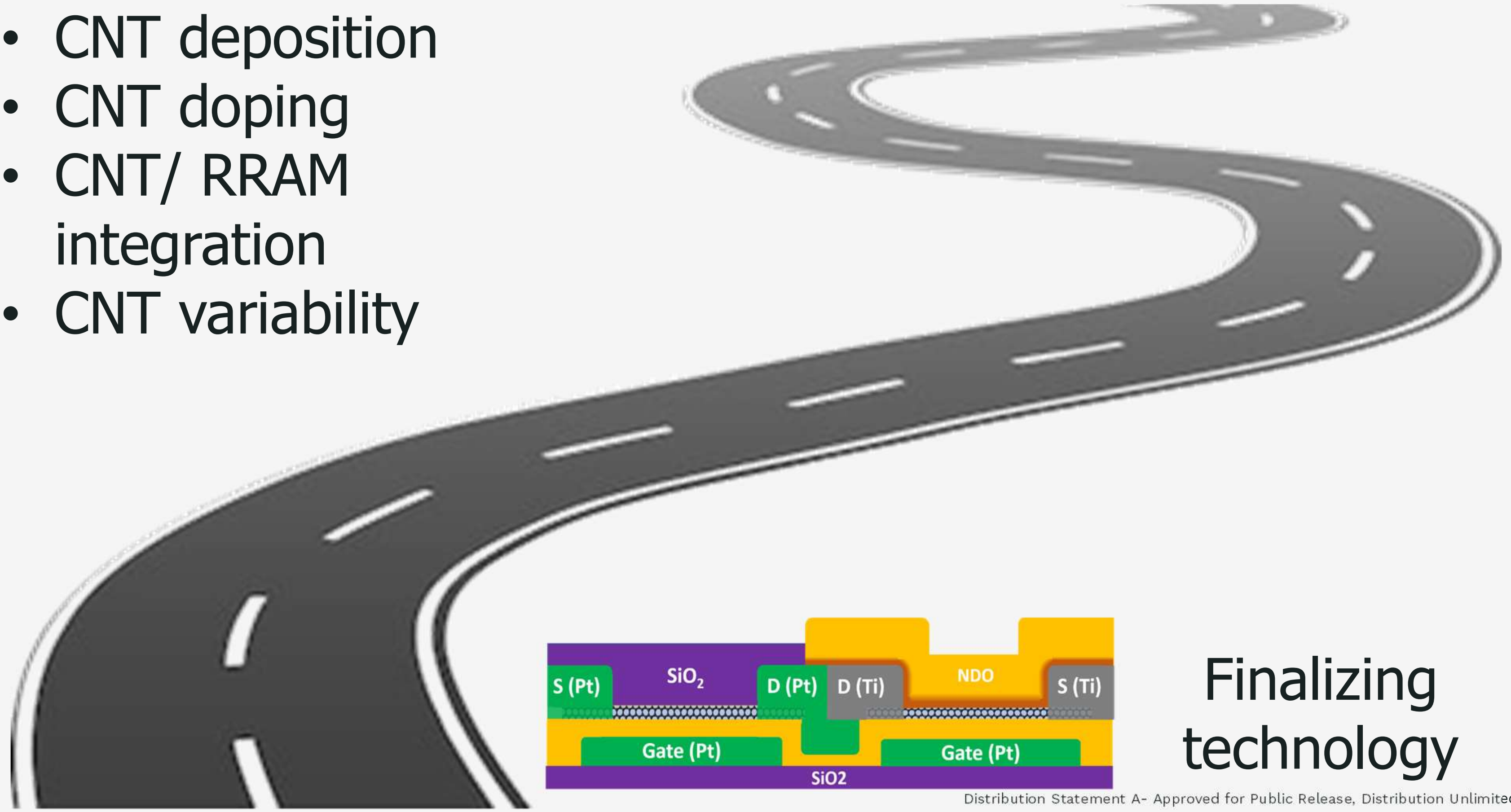
- Layer 4+: RRAM on-chip memory





ROADMAP

- CNT deposition
- CNT doping
- CNT/ RRAM integration
- CNT variability

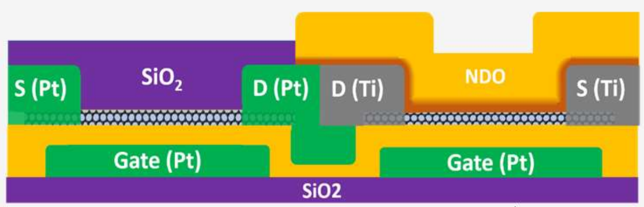


Finalizing technology

ROADMAP

- Technology transfer: CNT + RRAM modules
- 200 mm substrates

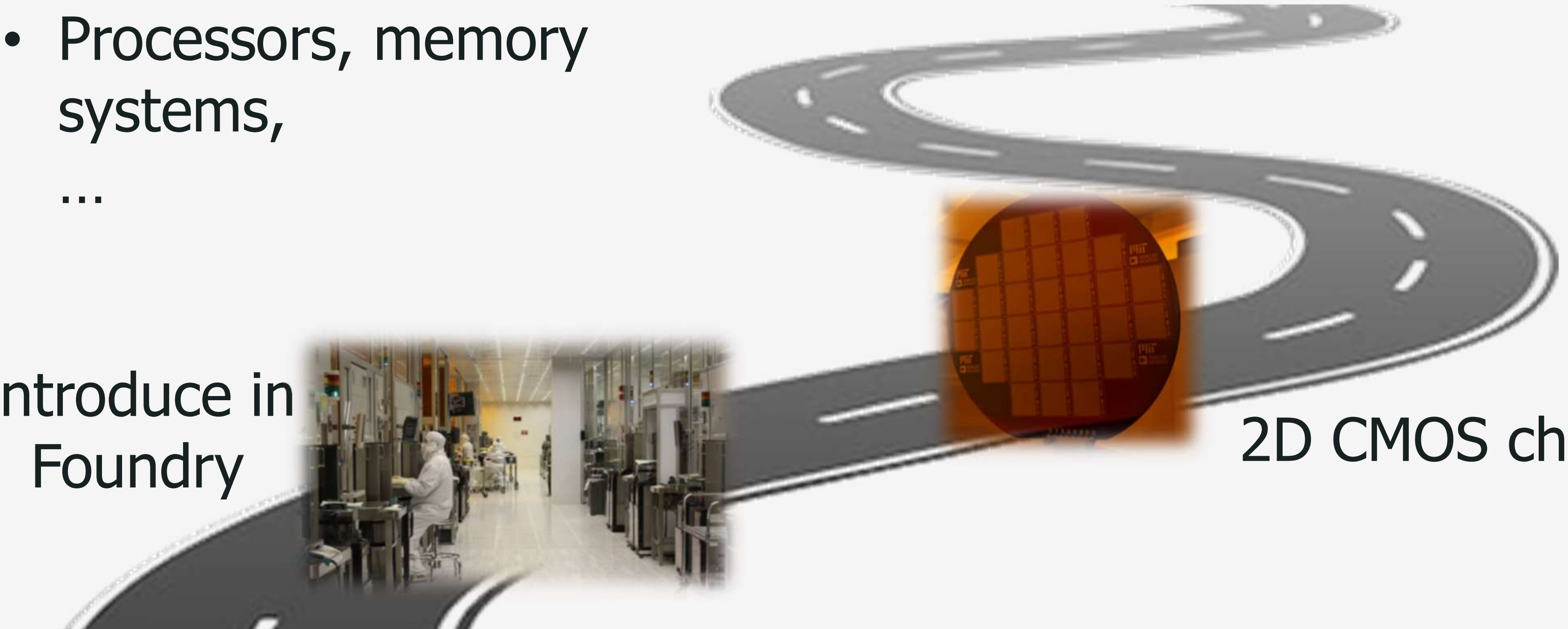
Introduce in Foundry



Finalizing technology

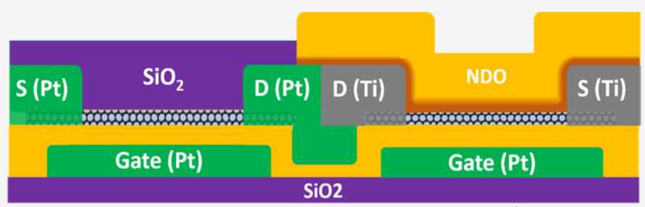
ROADMAP

- Processors, memory systems, ...



Introduce in Foundry

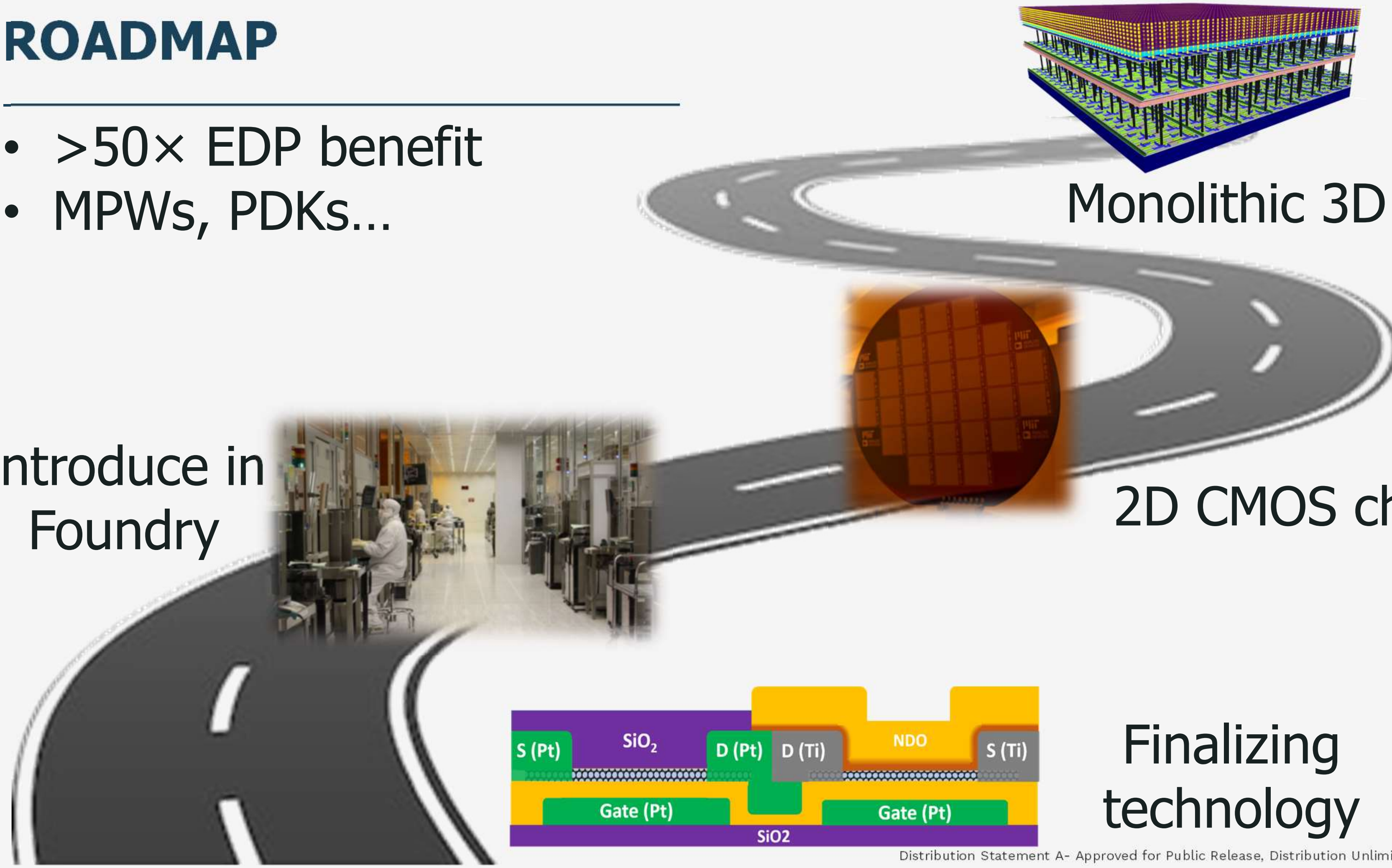
2D CMOS chips



Finalizing technology

ROADMAP

- >50× EDP benefit
- MPWs, PDKs…



TEAM


- CNFETs (optimize, transfer, PDK)
- monolithic 3D fabrication
- Program + system integration



- RRAM (optimize, transfer, PDK)
- monolithic 3D system design
- evaluation (benchmarking)


- stand-up CNFET, RRAM modules
- demo monolithic 3D ICs
- develop MPW offering

- improving CNT material
- high-volume CNT production

KEY TAKEAWAYS

- Monolithic heterogeneous 3D
  - Carbon nanotube FETs + Resistive RAM
- Target 50× performance benefits
  - Highly demanding abundant-data applications
- New beginning, broad technology foundation
  - More benefits through technology, architecture, software advances

Performance: energy × execution time

ERI ELECTRONICS RESURGENCE INITIATIVE

SUMMIT

2018 | SAN FRANCISCO, CA | **JULY 23-25**

SUBHASISH MITRA

STANFORD UNIVERSITY

3DSOC: DRAMATIC ADVANCES IN COMPUTING SYSTEMS

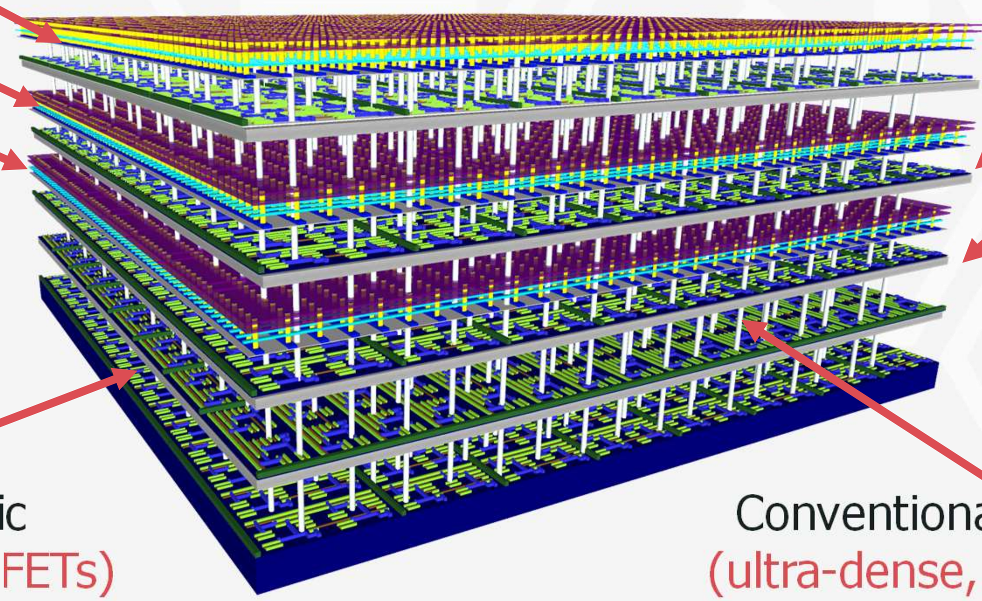
- This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA)
- The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government

OUR 3DSOC APPROACH

- Monolithic & heterogeneous 3D (inside silicon foundry)

Data storage (Resistive RAM)

Memory access
(Carbon nanotube FETs)

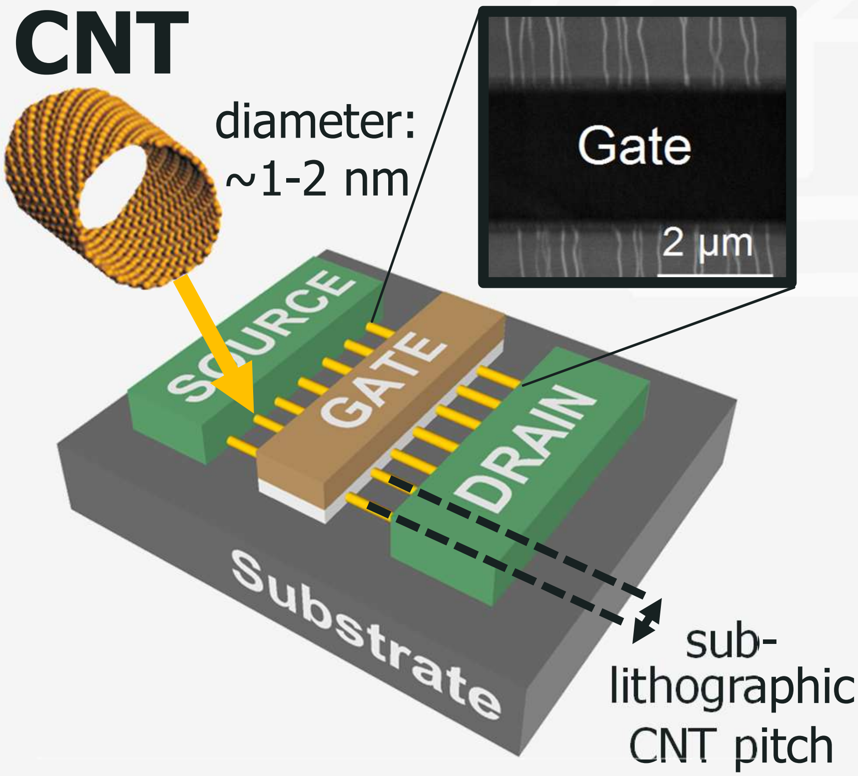


Computing Logic
(Carbon nanotube FETs)

Conventional metal vias
(ultra-dense, fine-grained)

COMPONENT TECHNOLOGY-LEVEL BENEFITS

CARBON NANOTUBE FETS (CNFETS)



Energy **D**elay **P**roduct

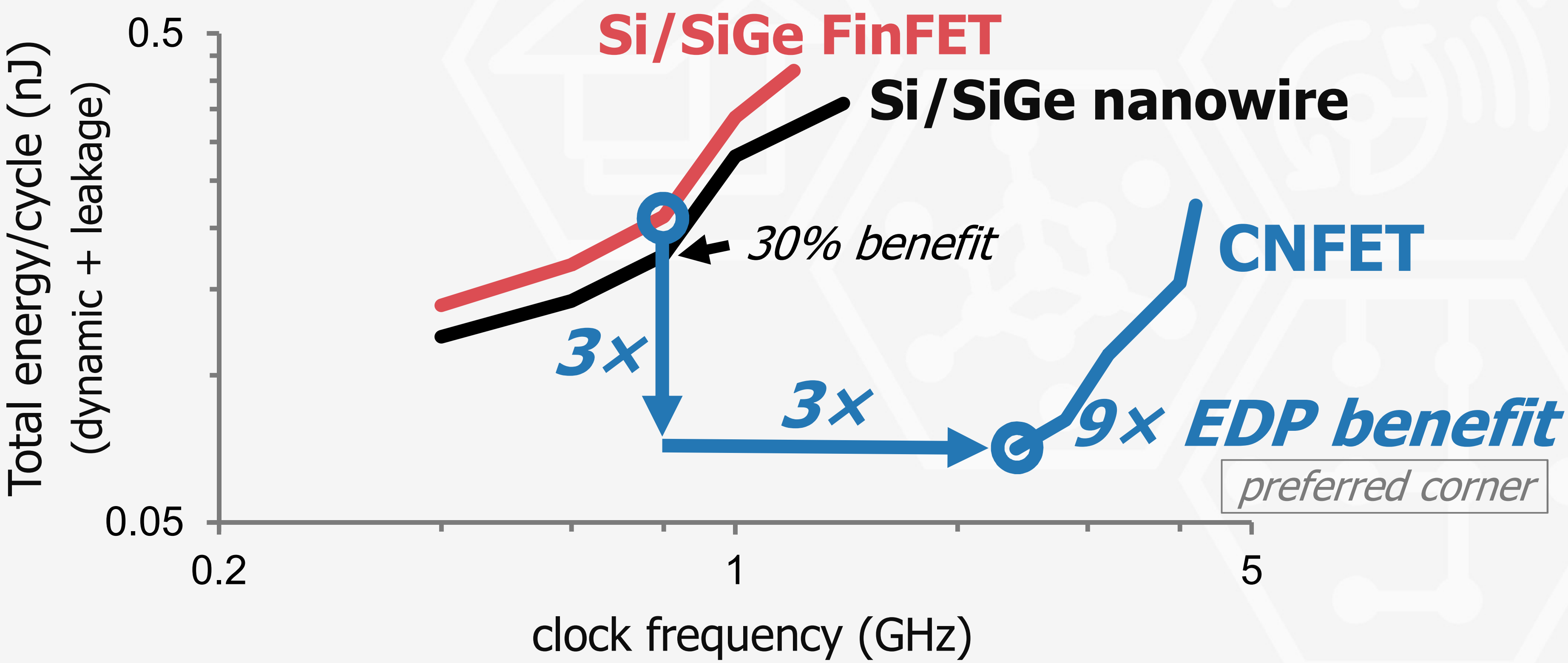
10× benefit

full-chip case studies

[IBM, IMEC, Stanford,

other commercial]

CNFET BENEFITS: OPENSPARC T2 PROCESSOR CORE



7 nm node, [Stanford + IMEC, Unpublished]

WHERE DO CNFET BENEFITS COME FROM?

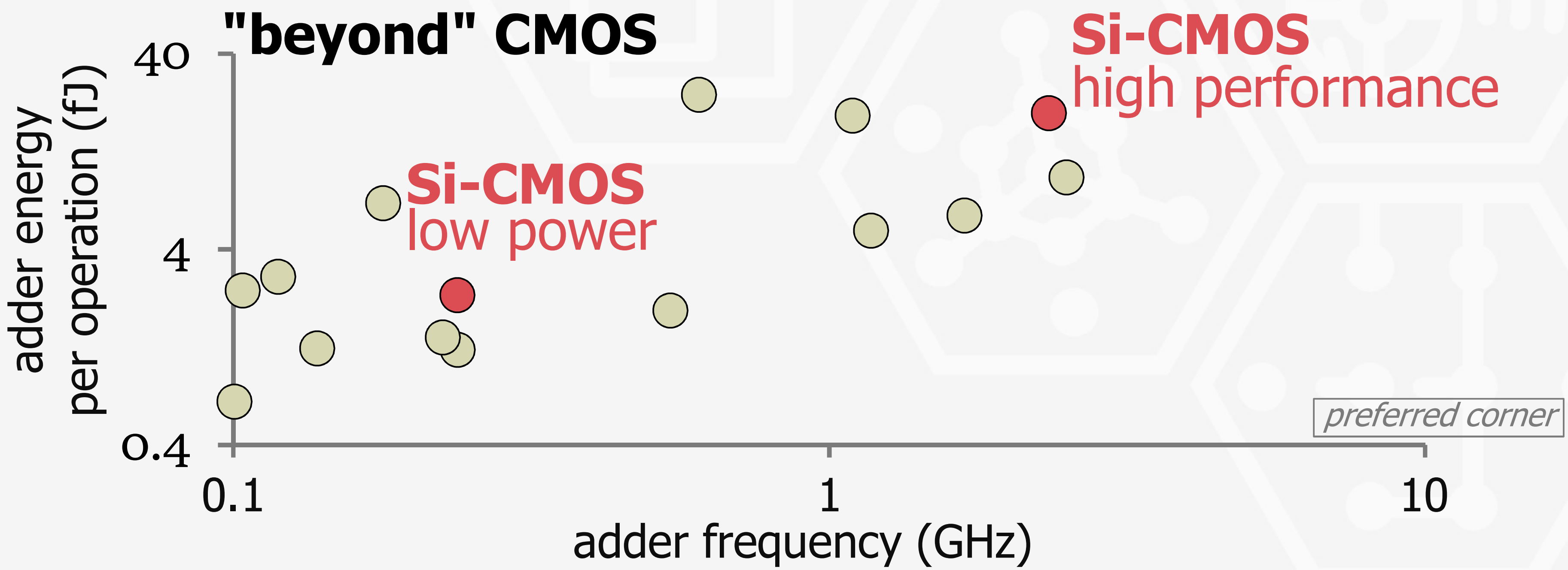
$$EDP \sim \underbrace{\boxed{CV^2}}_{\text{energy}} \times \underbrace{\boxed{CV/I}}_{\text{delay}}$$

- 20% lower V_{DD} — *$0.8^3 \sim 2\times$ benefit*
- 25% higher I_{ON} (same I_{OFF}) — *$\sim 1.25\times$ benefit*
- 2× lower circuit capacitance — *$2^2 \sim 4\times$ benefit*
  - Shorter L_G
  - Reduced parasitics
  - Smaller FET widths to meet timing

I_{OFF} = off-state leakage current

PUTTING INTO PERSPECTIVE

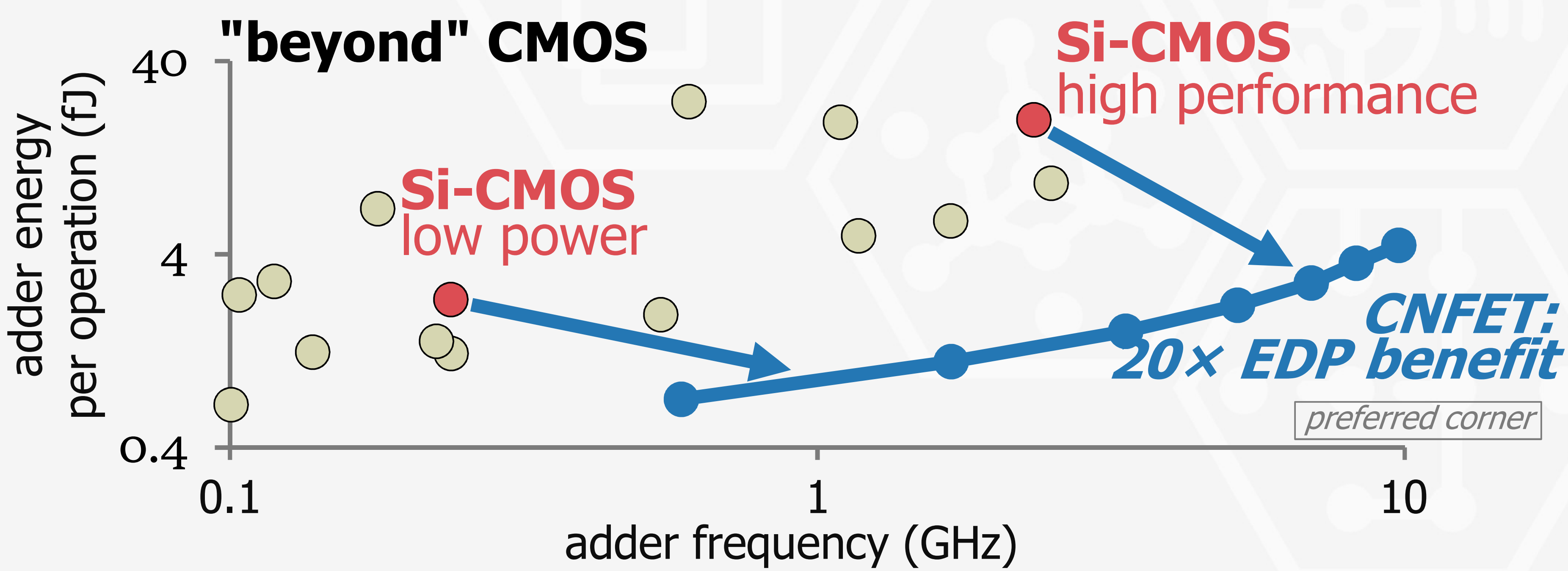
- Existing technology benchmarking



32-bit adder [Nikonov & Young, 2013 & 2015]

PUTTING INTO PERSPECTIVE

- Existing technology benchmarking **+ CNFETs**

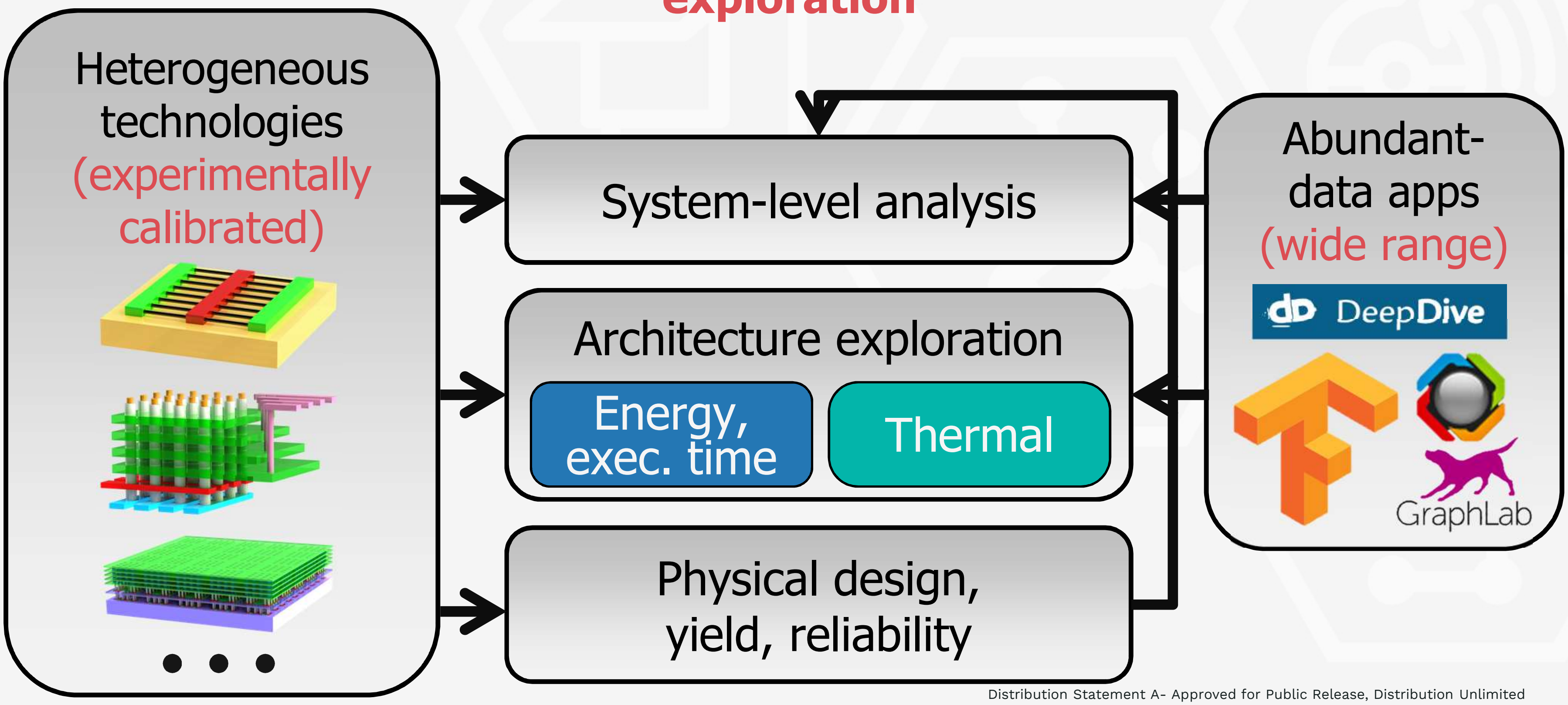


32-bit adder [Nikonov & Young, 2013 & 2015]

3DSOC-LEVEL BENEFITS

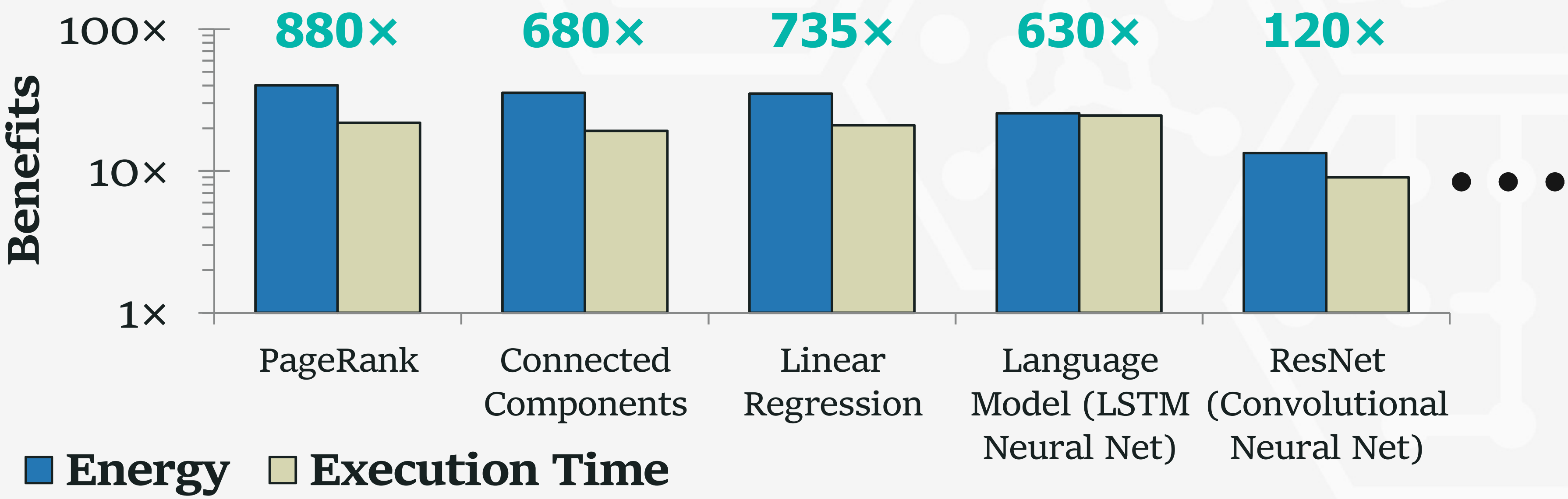
UNDERSTANDING SYSTEM-LEVEL BENEFITS

Simulation framework: technology, design & application co-exploration



3DSOC BENEFITS: DEEP LEARNING, GRAPH ANALYTICS, …

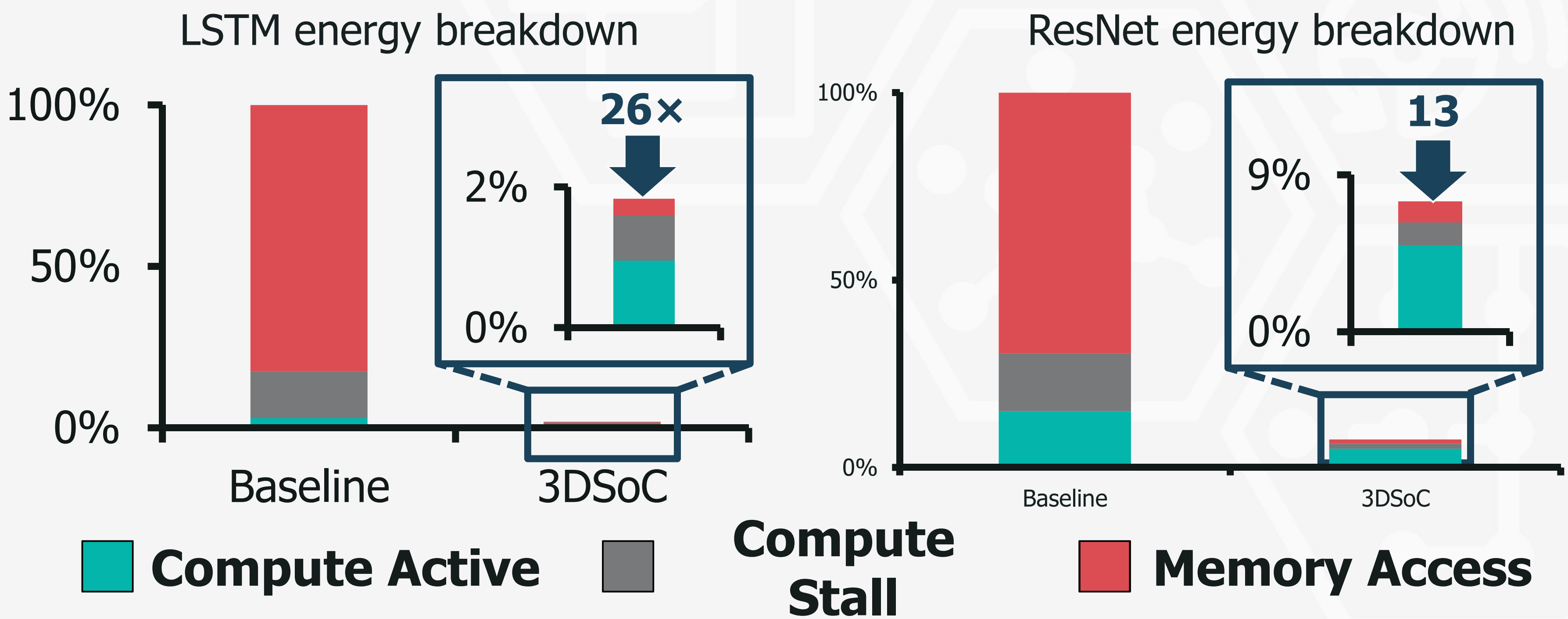
- Broad range of applications: ***~1,000× benefits*** (vs. baseline silicon 2D)
  - **Existing** software
- Silicon 2.5D: *2-4×* benefits



Baseline silicon 2D & 3DSoC: same technology node (e.g., 7nm)

UNDERSTANDING 3DSOC BENEFITS

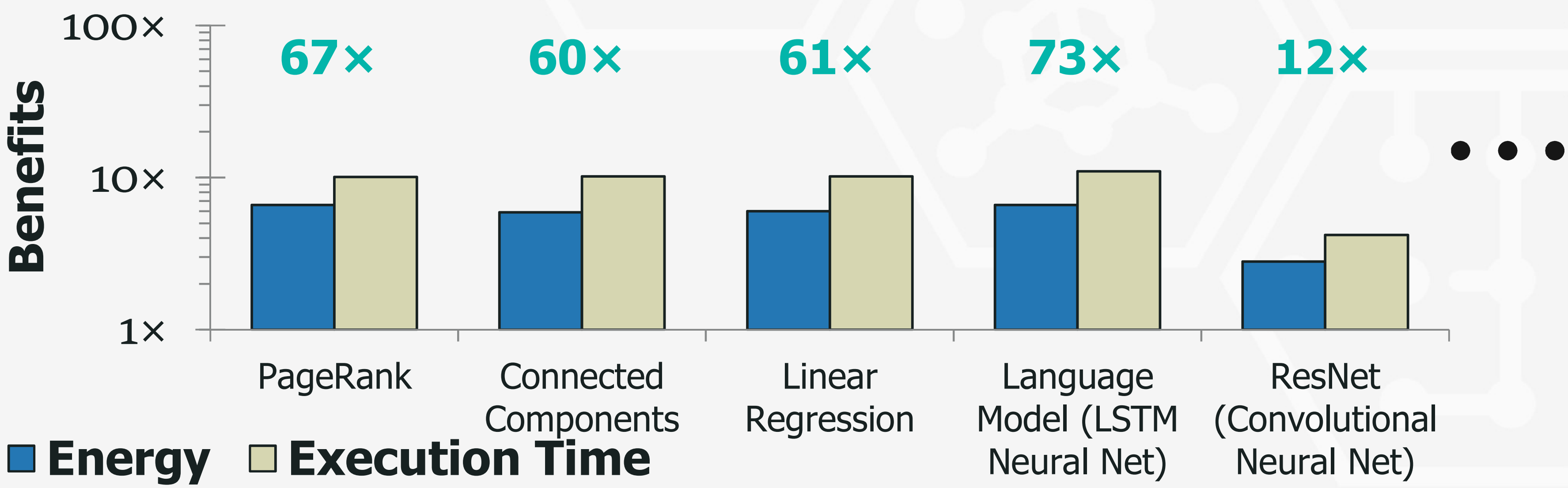
Ultra-dense & fine-grained vertical (3D) connectivity key



Baseline silicon 2D & 3DSoC: same technology node (e.g., 7nm)

3DSOC BENEFITS ACROSS TECHNOLOGY NODES

- **Dramatic benefits: broad range of applications**
  - 3D SoC (90nm node) vs.
  - Baseline silicon 2D (7nm node)



SPUR INNOVATION AT ALL LEVELS

- Dramatic benefits enable new application capabilities
  - New 3DSOC architectures not possible today
- Design enablement for new 3DSOC architectures
  - New 3DSOC PDKs, new 3DSOC EDA tools
- Reinvigorate electronic systems technology
  - Scalable path for several new generations of nanosystems

3DSOC: KEY TAKEAWAYS

- Monolithic heterogeneous 3D
  - Carbon nanotube FETs + Resistive RAM
- Target 50× performance benefits
  - Highly demanding abundant-data applications
- New beginning, broad technology foundation
  - Further benefits through technology, architecture, software advances

Performance: energy × execution time

ERI ELECTRONICS RESURGENCE INITIATIVE

SUMMIT

2018 | SAN FRANCISCO, CA | JULY 23-25